89

Reihe Ökonomie

Economics Series

DEA Problems under Geometrical or Probability Uncertainties of Sample Data

Karl S. Althaler, Tatjana Slavova

89

Reihe Ökonomie

Economics Series

DEA Problems under Geometrical or Probability Uncertainties of Sample Data

Karl S. Althaler, Tatjana Slavova

October 2000

Institut für Höhere Studien (IHS), Wien
Institute for Advanced Studies, Vienna

Contact:

Karl S. Althaler
☎: +43/1/599 91-227
email: althaler@ihs.ac.at

Tatjana Slavova
☎: +43/1/599 91-224
email: slavova@ihs.ac.at

Abstract

This paper discusses the theoretical and practical aspects of new methods for solving DEA problems under real-life geometrical uncertainty and probability uncertainty of sample data. The proposed minimax approach to solve problems with geometrical uncertainty of sample data involves an implementation of linear programming or minimax optimization, whereas the problems with probability uncertainty of sample data are solved through implementing of econometric and new stochastic optimization methods, using the stochastic frontier functions estimation.

Zusammenfassung

Diese Publikation behandelt die theoretischen und praktischen Aspekte der neuen Methoden zur Lösung von DEA-Problemen mit *real-life* geometrischer Unsicherheit und stochastischer Unsicherheit von Daten. Die vorgeschlagenen minimax-Methoden zur Lösung der geometrischen Unsicherheit von Daten beziehen die Implementierung der linearen Programmierung oder minimax-Optimierung mit ein, während die Probleme mit Unsicherheit der Wahrscheinlichkeit von Daten durch Implementierung von ökonometrischen und neuen stochastischen Optimierungsmethoden zur Schätzung der stochastischen Grenzfunktionen gelöst werden.

Keywords

DEA, sample data uncertainty, linear programming, minimax optimization, stochastic optimization, stochastic frontier functions

Schlagwörter

DEA, Ungewissheit von Daten, lineare Programmierung, minimax-Optimierung, stochastische Optimierungsmethoden, stochastische Grenzfunktionen

JEL Classifications

C81, D81, H72

Comments

We would like to thank our colleagues of the Institute for Advanced Studies in Vienna Alexander Schnabl and Raimund Alt for their helpful comments and Dimitar Dimitrov for the time he contributed to this paper.

Contents

1. Introduction

This paper is concerned with new methods for measuring the performance of firms (or "decision making units" (DMUs) which convert inputs into outputs. The methods of performance measurement that are proposed here may be used in many areas, including applications to private sector firms producing goods as well as to different service industries (such as travel agencies or restaurants), or to non-profit organizations (such as schools or hospitals). The methods may be used by a particular firm to analyze the relative performance of units within the firm.

These newly proposed methods differ according to the type of measures they produce, the data they require, and the assumptions they are based upon regarding the structure of production technology and the economic behavior of decision makers.

Data Envelopment Analysis (DEA) is an efficient technique for deciding on the relative efficiency of a DMU by comparing it with other DMUs engaged in making the same outputs from the same inputs. The DEA model uses a mathematical programming technique to estimate the efficiency frontier. This contrasts with the traditional econometric approach, which estimates an "average" relationship between inputs and outputs. As noted by Seiford and Thrall (1990), the econometric approach has a number of weaknesses. In order to estimate the coefficients of the production function, it requires the functional form to be pre-specified. It is possible only one output variable to take into account. The functional form will not, in general, be known, however, and adopting an arbitrary functional form will produce misspecification errors. It does not readily yield a summary judgment of efficiency, as only residuals are produced. The ability of the econometric model to identify sources of inefficiency is weak and influenced by outliers. Finally, by estimating a function on the basis of average response, it ignores the important distinction between firms which optimize their selection of inputs and those which do not.

By contrast, DEA is an extremal process, analyzes each firm separately and measures its relatively efficiency with respect to the entire set of DMUs being evaluated. It does not require any *a priori* assumption on the analytic form of production function. It is applicable to organizations characterized by multiple outputs and multiple inputs. The possibility to take into consideration multiple outputs is a special advantage of DEA compared to alternative methods, in particular to the traditional econometric approach. A DEA-based production model can also accommodate a variable that is neither an economic resource nor a product, such as attributes of the environment or the production process (e.g., Charnes et al., 1985). DEA provides solutions using standard techniques of linear programming, and thus provides the benefits of computational efficiency, dual variables and clear interpretations. The empirical orientation and absence of *a priori* assumptions have made it possible to measure efficiency from direct efficient frontier estimation in non-profit and regulated sectors as well as in profit-maximizing organizations. DEA both evaluates and identifies inefficiencies of DMUs and provides targets for improvement for inefficient DMUs. It can therefore also serve as a planning aid to management.

This efficiency question embraces both technical and scale efficiency. The former is concerned with efficiency in converting inputs to outputs (given the size of the DMU) and defined in terms of a production frontier as the ratio of potential and actual performance. Following Farrell (1957), the comparison of efficiency performance is made with the best of an industry, i.e. the observed industry standard. The efficiency frontier is constituted of those units, which are efficient relative to other units under evaluation. Efficiency computations are made relative to this frontier. The scale efficiency is concerned with whether the investigated DMU is operating at its optimal size (in comparison with other observed DMUs). This efficiency can be measured, when one believes that the technology is VRS (variable returns to scale). VRS occurs when a proportional increase in all inputs does not result in the same proportional increase in output. To obtain a scale efficiency measure for each firm, one can conduct both a CRS (constant return to scale) DEA and a VRS DEA. If there is a difference in the CRS and VRS technical efficiency scores for a particular firm, then this indicates that the firm has scale inefficiency.

In general, for DEA analysis one does not need weights of input and output variables (for example, price information). But if one is willing to consider a behavioral objective, such as cost minimization or revenue maximization, one can measure both technical and allocative efficiencies.

Usually, any estimation of an unknown production function of fully efficient firms with the use of sample data is based on the implementation of either a non-parametric piece-wise-linear DEA technology or a parametric function, such as the Cobb-Douglas form. Thus, any noise presented (e.g., due to measurement error, not accounting for environmental differences such as strikes, weather, etc.) may influence the shape and position of the piece-wise linear DEA frontier more than would be the case with the stochastic frontier approach. It means that stochastic frontiers are likely to be more appropriate than the piece-wise linear DEA frontier in the applications, where the sample data are heavily influenced by measurement errors, outliers, and environmental differences. However, in the non-profit service sector, where multiple-output production is important, random influences are less of an issue and prices are difficult to define, the DEA approach may often be the optimal choice. In this paper we will investigate this very important case of uncertainty.

Now we provide an outline of the contests of the subsequent four chapters in this paper. As noted above, we consider three different methodologies: DEA, stochastic frontiers and the Malmquist index. Each of these methods has one chapter devoted to it. For each method we first describe the basic methodology and provide a description of new investigations for DEA and stochastic frontiers. We will develop new DEA minimax methods for measuring DMU production efficiencies in cases of geometrically uncertain sample data and new stochastic frontier methods for measuring DMU production efficiencies in cases of geometrically uncertain sample data. In the last chapter we give details on the computer software used for efficiency estimation.

2. New DEA Minimax Methods

In this section we introduce and show the algorithms of applications of new DEA minimax methods under real-life uncertainties in input and/or output variables.

For DMU$_j$, $j = 1,2,...,n$ we use the notation $x_{qj}, q = 1,2,...,k,$ (or as vector x_j) for observation data on k inputs and $y_{rj}, r = 1,2,...,m,$ (or as vector y_j) for observation data on k outputs. The purpose of DEA is to construct a non-parametric envelope (or efficiency frontier) over the data points such that all observed points lie on or below the production frontier. In this paper we provide three models that are not yet investigated in the literature. The first model corresponds to the real-life output data inequality-type uncertainties:

$$y_{rj}^{\min} \le y_{rj} \le y_{rj}^{\max}, r = 1,2,...,m, j = 1,2,...,n \tag{1}$$

The second model is related to the input-output uncertainties of the simple inequality-type:

$$\begin{aligned} &x_{qj}^{\min} \le x_{qj} \le x_{qj}^{\max}, y_{rj}^{\min} \le y_{rj} \le y_{rj}^{\max}, \\ &q = 1,2,...,k, r = 1,2,...,m, j = 1,2,...,n \end{aligned} \tag{2}$$

The third model matches to the input-output uncertainties of the general type given by the inclusions

$$(y_j, x_j) \in \Omega_j, j = 1,2,...,n, \tag{3}$$

for certain sets Ω_j of input and output vectors $(y_j, x_j) = (y_{1j}, y_{2j},..., y_{m,j}, x_{1j}, x_{2j},..., x_{kj})$. For simplicity's sake, let us designate the production data of the decision making unit (y_i, x_i). by DMU$_i$

For each DMU$_i$ $, i = 1,2,...,n$, one would like to obtain a measure of the ratio of all outputs over all inputs, such as $u^{\mathrm{T}} y_i / n^T x_i$, where u is a vector of output weights and n is a vector of input weights. The optimal weights are calculated as the solution of the following mathematical programming problem:

$$\begin{aligned}
&\max_{u,n}(u^{T} y_i / n^{T} x_i),\\
&u^{T} y_j / n^{T} x_j \leq 1,\ j = 1,2,\ldots,n,\\
&u,n \geq 0
\end{aligned} \tag{4}$$

One knows that this problem has an infinite number of solutions. Indeed, if (u^*, n^*) is a solution, then $(cu^*, cn^*), c > 0$ is another solution. To avoid this, one can impose the constraint $n^T x_i = 1$, which provides:

$$\begin{aligned}
&\max_{u,n} u^{T} y_i,\\
&u^{T} y_j - n^{T} x_j \leq 0,\ j = 1,2,\ldots,n,\\
&v^{T} x_i = 1,\\
&u,n \geq 0
\end{aligned} \tag{5}$$

Using the duality in linear programming, one can derive an equivalent input-orientated envelopment form of this problem:

$$\begin{aligned}
&\min_{q,l} q,\\
&q x_i - \sum_{j=1}^{n} x_j l_j \geq 0,\\
&\sum_{j=1}^{n} y_j l_j - y_i \geq 0,\\
&l \geq 0,
\end{aligned} \tag{6}$$

where q is scalar and l is a $n \times 1$ vector of constants.

From the inequality $k + m < n + 1$ follows that the envelopment form (6) involves fewer constraints than the multiplier form (5), and generally is the preferred form to solve the problem. But the u and n weights can be interpreted as normalized shadow prices. For this purpose the multiplier form (5) is estimated in a number of studies. The obtained value of q satisfy $q^* \leq 1$. It is considered as the efficiency score for the DMU$_i$ .The value $q^* = 1$ indicates the point (x_i, y_i) on the frontier and hence a technically efficient DMU$_i$.

However, the piece-wise linear form of the non-parametric frontier in DEA can cause difficulties in efficiency measurement because of the sections of the piece-wise linear frontier that may run parallel to the axes. In this case, one could reduce the amount of input used and still produce the same output (this is known as *input slack* in the literature), or one could increase the amount of output produced and still use the same input (this is known as *output* *slack*). Thus it could be argued that both the measure of technical efficiency q^* and any non-zero input or output slacks should be reported to provide an accurate indication of the technical efficiency of a DMU in a DEA analysis. Taking into account that for the *i*-th DMU the output slacks will be equal to zero only if $\sum_{j=1}^{n} y_j l_j - y_i = 0$ and, as well, the input slacks will be equal to zero only if $q^* x_i - \sum_{j=1}^{n} x_j l_j = 0$ (for the given optimal values of q^* and l), it was suggested (e.g. Coelli, 1996b) to consider a second-stage linear programming problem (7) in order to move to an efficient frontier point by maximizing the sum of slacks required to move from an inefficient frontier point to an efficient frontier point:

$$\begin{aligned}
&\max_{s^+, s^-, l} (e^{+T} s^+ + e^{-T} s^-), \\
&\sum_{j=1}^{n} x_j l_j + s^+ = q^* x_i, \\
&\sum_{j=1}^{n} y_j l_j - s^- = y_i, \\
&s^+, s^- \geq 0, l \geq 0,
\end{aligned} \tag{7}$$

where e^+ and e^- are $k \times 1$ and $m \times 1$ vectors of ones, respectively, s^+ is a $k \times 1$ vector of input slacks, and s^- is a $m \times 1$ vector of output slacks. It should be taken into account that in this second-stage linear programming problem (7), the q^* is not a variable. Its value is taken from the first-stage results. This second-stage linear program must be solved for each DMU$_i$ of the n DMUs involved. The second major problem associated with the above second-stage approach is that it is not invariant to units of measurement. Thus, the alteration of the units of measurement (say for a fertilizer input from kilograms to tons, while leaving other units of measurement unchanged) could result in the identification of different efficient boundary points and could hence result in different s^+, s^- and l. As a result of this problem, many studies simply solve the first-stage problem (3) to calculate the efficiency score q^* for each DMU and ignore the slacks completely, or they report both the efficiency score q^* and the residual slacks $s^+ = q^* x_i - \sum_{j=1}^{n} x_j l_j, s^- = -y_i + \sum_{j=1}^{n} y_j l_j$. Because of this, there are three main choices regarding the treatment of slacks:

1. One-stage DEA, in which the problem (6) is solved to calculate the efficiency score q^* for each DMU and to calculate slacks residually,

$$s^+ = q^* x_i - \sum_{j=1}^{n} x_j l_j, s^- = -y_i + \sum_{j=1}^{n} y_j l_j;$$

2. Two-stage DEA, in which the first-stage problem (6) is solved to calculate the efficiency score q^* for each DMU, and the second-stage problem (7) is solved with given q^* to move from an inefficient frontier point to an efficient frontier point;

3. Multi-stage DEA, where one conducts a sequence of radial LP's to identify the efficient projected points.

From the given DEA data, one can derive four different production possibility sets

$$T = \left\{ (x, y) : \sum_{j=1}^{n} x_j l_j \leq x, \sum_{j=1}^{n} y_j l_j \geq y, l \in \Gamma \right\}$$

under four different sets Γ:

$$\Gamma_{CRS} = \{ l : l \geq 0 \},$$

$$\Gamma_{VRS} = \left\{ l : \sum_{j=1}^{n} l_j = 1, l \geq 0 \right\},$$

$$\Gamma_{NIRS} = \left\{ l : \sum_{j=1}^{n} l_j \leq 1, l \geq 0 \right\},$$

$$\Gamma_{NDRS} = \left\{ l : \sum_{j=1}^{n} l_j \geq 1, l \geq 0 \right\}$$

Thus, one obtains four production possibility sets: T_{CRS} is T with Γ_{CRS}, T_{VRS} is T with Γ_{VRS}, T_{NIRS} is T with Γ_{NIRS}, T_{NDRS} is T with Γ_{NDRS}. The DMU$_i$ efficiencies q may be evaluated as solutions of the following four input-oriented LP models (in all four LPs we minimize the multiple q of DMU$_i$'s inputs required to produce at least DMU$_i$'s outputs, minus a small multiple e of the sum of slacks on each input and output (e.g. Ali and Seiford, 1993):

$$
\begin{aligned}
&\min_{s^+,s^-,q} q - e(e^{+T}s^+ + e^{-T}s^-), \\
&\sum_{j=1}^{n} x_j l_j + s^+ = qx_i, \\
&\sum_{j=1}^{n} y_j l_j - s^- = y_i, \\
&s^+, s^- \geq 0,\ l \geq \Gamma
\end{aligned}
\tag{8}
$$

Here e^+ and e^- are $k \times 1$ and $m \times 1$ vectors of 1's, respectively, and Γ is one of the four different sets: Γ_{CRS}, Γ_{VRS}, Γ_{NIRS} or Γ_{NDRS}.

In the cases of the uncertainties (1) and (2), the LP (8) take a form of the following four optimization problems (9), (10), (11) and (12):

$$
\begin{aligned}
&\min_{s^+,s^-,q} q - e(e^{+T}s^+ + e^{-T}s^-), \\
&\sum_{j=1}^{n} x_j l_j + s^+ = qx_i, \\
&\sum_{j=1}^{n} y_j^{\max} l_j - s^- = y_i^{\min}, \\
&s^+, s^- \geq 0,\ l \geq \Gamma
\end{aligned}
\tag{9}
$$

$$
\begin{aligned}
&\min_{s^+,s^-,q} q - e(e^{+T}s^+ + e^{-T}s^-), \\
&\sum_{j=1}^{n} x_j^{\min} l_j + s^+ = qx_i^{\max}, \\
&\sum_{j=1}^{n} y_j^{\max} l_j - s^- = y_i^{\min}, \\
&s^+, s^- \geq 0,\ l \geq \Gamma
\end{aligned}
\tag{10}
$$

$$
\begin{aligned}
&\min_{s^+,s^-,q} q - e(e^{+T}s^+ + e^{-T}s^-), \\
&\sum_{j=1}^{n} x_j l_j + s^+ = qx_i, \\
&\sum_{j=1}^{n} y_j^{\min} l_j - s^- = y_i^{\max}, \\
&s^+, s^- \geq 0,\ l \geq \Gamma
\end{aligned}
\tag{11}
$$

$$\begin{aligned}
&\min_{s^+,s^-,q} q - e(e^{+T}s^+ + e^{-T}s^-),\\
&\sum_{j=1}^{n} x_j^{\max} l_j + s^+ = q x_i^{\min},\\
&\sum_{j=1}^{n} y_j^{\min} l_j - s^- = y_i^{\max},\\
&s^+, s^- \geq 0,\ l \geq \Gamma
\end{aligned} \tag{12}$$

Let q^1, q^2, q^3 and q^4 be the solutions of (9), (10), (11), and (12). Then q^1 and q^3 are the greatest lower and upper boundaries of the DMU*i*'s efficiency q under uncertainties (1), i.e. $q \in [q^1, q^3]$, and q^2 and q^4 are the greatest lower and upper boundaries of the DMU*i*'s efficiency q under uncertainties (2), $q \in [q^2, q^4]$. Thus, a measure m_1 of the DMU*i*'s efficiency uncertainty in relation to the uncertainty (1) is evaluated by the difference $q^3 - q^1$, and a measure m_2 of the DMU*i*'s efficiency uncertainty in relation to the uncertainty (2) is evaluated by the difference $q^4 - q^2$.

In order to evaluate the greatest lower boundary q^5, the upper boundary q^6 and the measure $m_3 = q^6 - q^5$ of the DMU*i*'s efficiency uncertainty in relation to the uncertainty (3), we will consider the following three types of the uncertainty set Ω_j:

(i) a finite set $\Omega_j = \{(y_j^1, x_j^1), (y_j^2, x_j^2), \ldots, (y_j^{n_j}, x_j^{n_j})\}$,

(ii) a convex hull $\Omega_j = con\{(y_j^1, x_j^1), (y_j^2, x_j^2), \ldots, (y_j^{n_j}, x_j^{n_j})\}$,

i.e. $$\Omega_j = \left\{(y,x) : (y,x) = \sum_{l=1}^{n_j} m_l (y_j^l, x_j^l), \sum_{l=1}^{n_j} m_l = 1, m_l \geq 0\right\},$$

and

(iii) a polytope $\Omega_j = \{(y_j, x_j) : A_j y_j + B_j x_j \leq b_j\}$,

with some given $p \times m$ matrix A_j, given $p \times k$ matrix B_j and given $p \times 1$ vector b_j.

To evaluate q^5 and q^6 in the cases (i) and (ii) one has to do the following:

1) replace all the products $x_j l_j$ and $y_j l_j$ in the model (8) by $\sum_{l=1}^{n_j} x_j^l l_j^l$ and $\sum_{l=1}^{n_j} y_j^l l_j^l$ for $j \neq i$;

2) calculate solutions q_l of the augmented model (8) for each $(x_i^l, y_i^l), l = 1,\ldots,n_i$;

3) evaluate $q^5 = \min\limits_{l=1,\ldots,n_i} q_l, q^6 = \max\limits_{l=1,\ldots,n_i} q_l$.

To estimate q^5 in the general case (iii), one has to solve the following optimization problem:

$$
\begin{aligned}
&\min_{x_i, y_i} \min_{s^+, s^-, q} q - \boldsymbol{e}(e^{+T} s^+ + e^{-T} s^-),\\
&\sum_{j=1}^{n} x_j l_j + s^+ = q x_i \ ,\\
&\sum_{j=1}^{n} y_j l_j - s^- = y_i \ ,\\
&A_j y_j + B_j x_j \leq b_j, \ j = 1,\ldots,n,\\
&s^+, s^- \geq 0, \ l \geq \Gamma
\end{aligned}
\tag{13}
$$

And to evaluate q^6 in the case (iii), one has to solve the following optimization problem:

$$
\begin{aligned}
&\max_{x_i, y_i} \min_{s^+, s^-, q} q - \boldsymbol{e}(e^{+T} s^+ + e^{-T} s^-),\\
&\sum_{j=1}^{n} x_j l_j + s^+ = q x_i \ ,\\
&\sum_{j=1}^{n} y_j l_j - s^- = y_i \ ,\\
&A_j y_j + B_j x_j \leq b_j, \ j = 1,\ldots,n,\\
&s^+, s^- \geq 0, \ l \geq \Gamma
\end{aligned}
\tag{14}
$$

3. New Stochastic Frontier Methods

As noted by Yunos and Hawdon (1997), the major limitations of the DEA methods relate to the treatment of uncertainty. To the extent that there are errors of measurement, there will be uncertainty surrounding the efficiency calculations. Some progress has been made towards introducing uncertainty into DEA models, but as yet no generally agreed methods exist regarding its treatment.

In this section we will propose new stochastic frontier methods for estimating frontier functions and also for measuring the efficiency of production in cases of probability uncertainties.

3.1. Stochastic Frontier Production Functions

The stochastic frontier production function was independently proposed by Aigner, Lovell and Schmidt (1977) and Meeusen and van den Broeck (1977). The original specification involves a production function specified for cross-sectional data, which had a two-component error term, one to account for random effects and another to account for technical inefficiency. This model can be expressed in the following form. Assume that the production function of fully efficient DMUs is known in Cobb-Douglas form

$$\ln(y_i) = x_i \boldsymbol{b} - U_i, i = 1,2,\ldots,n,$$

where $\ln(y_i)$ is the logarithm of the (scalar) output for the DMU$_i$; x_i is a $1\times(k+1)$ input vector, whose first elements is “1” and the remaining elements are the logarithms of the k-input quantities used by the DMU$_i$; $\boldsymbol{b}$ is a $(k+1)\times 1$ vector of unknown parameters to be estimated, and U_i is a non-negative random variable, associated with technical inefficiency in the production of DMUs in the industry involved. In this case, the technical efficiency TE_i of the DMU$_i$ firm may be evaluated as

$$TE_i = \frac{y_i}{e^{x_i \boldsymbol{b}}} = \frac{e^{x_i \boldsymbol{b} - U_i}}{e^{x_i \boldsymbol{b}}} = e^{-U_i}$$

In case of probability uncertainties, the value y_i is obtained with a random error V_i, i.e., one has the *stochastic* frontier production function

$$\ln(y_i) = x_i \boldsymbol{b} - U_i + V_i, i = 1,2,\ldots,n,$$

Assume that $V_i s$ are i.i.d. truncations at zero of a $N(m, s^2)$ random variable, i.e., its distribution $p^+(x)$ is defined as $p^+(x) = 2p(x)$ for $x \geq 0$, and $p^+(x) = 0$ for $x < 0$, where

$$p(x) = \frac{1}{\sqrt{2p}\, s} e^{-\frac{(x-m)^2}{2s^2}}$$

is the normal distribution. Then, in the case of the $V_i s$ being i.i.d. $N(0, s_v^2)$, Battese and Corra (1977) showed that the log-likelihood function $\ln(L)$ is evaluated as

$$2\ln(L) = -n\ln(\frac{p}{2}) - n\ln(s_s^2) + 2\sum_{i=1}^{n} \ln(1-\Phi(z_i)) - \frac{\sum_{i=1}^{n}(\ln(y_i) - x_i b)^2}{s_s^2}$$

where $z_i = \sqrt{\frac{g}{1-g}} \frac{(\ln(y_i) - x_i b)}{s_s}, g = \frac{s^2}{s_s^2}, s_s^2 = s^2 + s_n^2,$ and $\Phi(\cdot)$ is the cumulative

distribution function of the standard normal random variable. The maximum-likelihood estimates of b, s_s^2 and g can be obtained by finding the maximum of the $\ln(L)$. And the mathematical expectation (mean) of the technical efficiency can be calculated as

$$E(e^{-U_i}) = 2(1-\Phi(\sqrt{g}\, s_s))e^{-\frac{g s_s^2}{2}}$$

The best predictor for U_i is the conditional expectation of U_i, given the value $e_i = V_i - U_i = \ln(y_i) - x_i b$. It was obtained by Jondrow, Lowell, Materov and Schmidt (1982) as

$$E(U_i | e_i) = -g e_i + s_A \frac{f(\frac{g e_i}{s_A})}{(1-\Phi(\frac{g e_i}{s_A}))},$$

where $s_A = \sqrt{g(1-g)}\, s_s$ and f is the density function of a standard normal random variable. Battese and Coelli (1988) point out that the best predictor of $\exp(-U_i)$ is obtained by using

$$E(e^{-U_i} \left| e_i \right.) = \frac{1 - \Phi(s_A + g e_i / s_A)}{1 - \Phi(g e_i / s_A)} \exp(g e_i + \frac{s_A^2}{2})$$

This stochastic model has been used in a vast number of empirical applications over the past two decades. The model specification has also been altered and extended in a number of ways to cover more general distributional assumptions for the U_i and, furthermore, the consideration of panel data and time-varying technical efficiencies.

3.2. Panel Data Models and Time-varying Technical Efficiencies

If a number of DMUs are observed over a number of time periods $t = 1,2,...,T$, the data obtained are known as *panel data.* Battese and Coelli (1992) propose a stochastic frontier production function for panel data which has firm effects, assumed to be distributed as truncated normal random variables that may vary systematically with time. The model may be expressed as:

$$Y_{it} = x_{it} b + (V_{it} - U_{it}), i = 1,2,...,n, t = 1,...,T, \tag{15}$$

where Y_{it} is (the logarithm of) the production of the DMU*i* in the t-th time period; x_{it} is a $1 \times (k+1)$ vector, whose first element is "1" and the remaining elements are the logarithms of the k-input quantities used by the DMU*i* in the t-th time period; b is a $(k+1) \times 1$ vector of unknown parameters to be estimated; V_{it} are random variables, assumed to be i.i.d. $N(0, s_v^2)$ and independent of the $U_{it} = U_i \exp(-h(t-T))$. U_i are non-negative random variables associated with technical inefficiency in production. They are assumed to be i.i.d. truncated at zero of the $N(m, s_U^2)$ distribution. h is the parameter to be estimated, and the panel of data need not be complete (i.e. unbalanced panel data).

This model formulation includes a number of the special cases that have appeared in the literature. Setting h to be zero provides the time-invariant model set out in Battese, Coelli and Colby (1989). The additional restriction of m equal to zero reduces the model to Model One in Pitt and Lee (1981). Setting $T = 1$, one returns to the original cross-sectional, half-normal formulation of Aigner, Lovell and Schmidt (1977). If the cost function estimation is selected, one can estimate the model specification in Hughes (1988) and the Schmidt and Lovell (1979) specification, which assumed allocative efficiency. The latter two specifications are the cost function analogues of the production functions in Battese and Coelli (1988) and Aigner, Lovell and Schmidt (1977), respectively. A number of empirical studies (Pitt and Lee, 1981) estimate stochastic frontiers and predict firm-level efficiencies by using these estimated functions. Then they regressed the predicted efficiencies upon firm-specific variables (such as managerial experience, ownership characteristics, etc.) in an attempt to identify some of the reasons for differences in predicted efficiencies between firms in an industry. Kumbhakar, Ghosh and McGukin (1991) and Reifschneider and Stevenson (1991) proposed stochastic frontier models in which the inefficiency effects U_i are expressed as an explicit function of a vector of firm-specific variables and a random error. Battese and Coelli (1995) develop the Kumbhakar, Ghosh and McGukin (1991) model specification to be expressed as:

$$Y_{it} = x_{it} b + (V_{it} - U_{it}), i = 1,2,...,n, t = 1,...,T, \tag{16}$$

where Y_{it}, x_{it} and b are as defined earlier; the V_{it} are random variables, assumed to be i.i.d. $N(0, s_v^2)$, and independent of the non-negative random variables U_{it}, associated with technical inefficiency in production. U_{it} are assumed to be independently distributed as truncations at zero of the $N(m_{it}, s_U^2)$ distribution, where:

$$m_{it} = z_{it} d, \tag{17}$$

z_{it} is a $p \times 1$ vector of variables which may influence the efficiency of a firm and d is an $1 \times p$ vector of the parameters to be estimated.

If $T = 1$ and z_{it} contains the value one, then this model is reduced to the truncated normal specification (Stevenson, 1980), with d having the same interpretation as the parameter m has. The log-likelihood function of this model is presented in the appendix of the paper of Battese and Coelli (1993).

3.3 New Non-linear Model of the Stochastic Frontier Production Function

We propose a new general non-linear model specification to be expressed as:

$$Y_{it} = f(x_{it}, b_1) - g(z_{it}, b_2) + V_{it}, g(z_{it}, b_2) \geq 0, i = 1,...,n, t = 1,...,T, \tag{18}$$

Y_{it} and x_{it} were defined earlier. z_{it} is a $p \times 1$ vector of variables which may influence the efficiency of a firm. $g(z_{it}, b_2)$ is a function which is assumed to account for technical inefficiency in production. V_{it} are random variables with parametric distribution functions $p_{it}(v, b_3)$; and $b = (b_1, b_2, b_3)$ is a vector of parameters to be estimated.

The maximum-likelihood estimates of b are obtained by finding the maximum of the likelihood function $L(b),$ under restrictions

$$g(z_{it}, b_2) \geq 0, i = 1,...,n, t = 1,...,T, \tag{19}$$

where $L(b)$ is defined as

$$L(b) = \prod_{i=1}^{n} \prod_{t=1}^{T} p_{ij}(Y_{ij} - f(x_{it}, b_1) + g(z_{it}, b_2), b_3), \tag{20}$$

To calculate the optimal solution b^*,

$$b^* = \arg\max_{b} L(b) : g(z_{it}, b_2) \geq 0, i = 1,..., n, t = 1,..., T,$$

we solve the following simpler optimization problem

$$b^* = \arg\max_{b} \sum_{i=1}^{n} \sum_{t=1}^{T} \ln p_{ij}(Y_{ij} - f(x_{it}, b_1) + g(z_{it}, b_2), b_3) :$$
$$g(z_{it}, b_2) \geq 0, i = 1,.., n, t = 1,..., T,$$

using numerical optimization algorithms of the [sub]gradient type (e.g. Beyko et al., 1983):

$$b^{k+1} = b^k + l_k w^k,$$
$$l_k > 0, \sum_{k=1}^{\infty} l_k = \infty, \lim_{k \to \infty} l_k = 0,$$

where

$$w_k = \nabla_b L(b^k) \text{ in case of } F(b_2) \geq 0, F(b_2) = \min_{it} g(z_{it}, b_2),$$

and

$$w_k = \nabla_b F(b^k) \text{ in case of } F(b_2) < 0.$$

One should note that the general model (19), (20) includes the Battese and Coelli (1995) model

$$Y_{it} = x_{it} b_1 - U_{it} + V_{it}, i = 1,2,..., n, t = 1,..., T, \tag{21}$$

where V_{it} are random variables, assumed to be i.i.d. $N(0, s_v^2)$; U_{it} are non-negative random variables, independently distributed as truncations at zero of the $N(z_{it} b_2, s_U^2)$ distribution; and $b = (b_1, b_2)$ is a vector of the parameters to be estimated.

3.4. Cost Functions

All of the above specifications have been expressed in terms of a production function, with the U_i interpreted as technical inefficiency effects which cause the firm to operate below the stochastic production frontier. To specify a stochastic frontier *cost function*, we simply alter the error term specification from $V_i - U_i$ to $V_i + U_i$. This substitution would transform the production function

$$\ln(y_i) = x_i b - U_i, i = 1,2,...,n,$$

into the Schmidt and Lovell (1979) cost function:

$$Y_i = x_i b + (V_i + U_i), i = 1,2,...,n, \qquad (22)$$

where Y_i is the (logarithm of the) cost of production of the DMU$_i$; x_i is a $k \times 1$ vector of (transformations of the) input prices and output of the i-th firm; b is an vector of unknown parameters; the V_i are random variables, assumed to be i.i.d. $N(0, s_v^2)$ and independent of the non-negative random i.i.d. $\left|N(0, s_U^2)\right|$ variables U_i. The variables U_i account for the cost of inefficiency in production.

In this cost function, U_i now defines how far the firm operates above the cost frontier. If allocative efficiency is assumed, U_i is closely related to the cost of technical inefficiency. If this assumption is not made, the interpretation of U_i in a cost function is less clear, with both technical and allocative inefficiencies possibly involved. The exact interpretation of these cost efficiencies will depend upon the particular application. The log-likelihood function of this model presented in the appendix of the paper of Schmidt and Lovell (1979) is the same as that of the production frontier, except for a few sign changes.

4. Efficiency Predictions under Uncertainties

In this section we will represent some new methods of efficiency predictions.

4.1. The Malmquist Total Factor Productivity (TFP) Index

The Malmquist index is defined through input distance functions and output distance functions that allow us to describe a multi-input, multi-output production technology without the need to specify a behavioral objective (such as cost minimization or profit maximization). The output distance function is defined on the input set $P(x)$,

$$P(x) = \{y : x \text{ can produce } y\},$$

as

$$d_0(x, y) = \min\{d : (y/d) \in P(x)\} \tag{23}$$

The Malmquist TFP index $m_0(y_s, x_s, y_t, x_t)$ measures the TFP change between period s (the base period) and period t (Färe et al. (1994)) by

$$m_0(y_s, x_s, y_t, x_t) = \left[\frac{d_0^s(y_t, x_t)}{d_0^s(y_s, x_s)} \frac{d_0^t(y_t, x_t)}{d_0^t(y_s, x_s)}\right]^{1/2}, \tag{24}$$

where the notation $d_0^s(y_t, x_t)$ represents the distance from the period t observation to the period s technology. Thus, $m_0(y_s, x_s, y_t, x_t)$ measures the TFP change between two data points by calculating the ratio of the distances of each data point relative to a common technology.

An equivalent way of writing this productivity index is

$$m_0(y_s, x_s, y_t, x_t) = \frac{d_0^t(y_t, x_t)}{d_0^s(y_s, x_s)} \left[\frac{d_0^s(y_t, x_t)}{d_0^t(y_t, x_t)} \frac{d_0^s(y_s, x_{ts})}{d_0^t(y_s, x_s)}\right]^{1/2}, \tag{25}$$

That is, the efficiency change is equivalent to the ratio of the Farrell technical efficiency in period t to the Farrell technical efficiency in period s. The remaining part of the index in equation (25) is a measure of technical change. Thus the two terms in equation (25) are:

$$\text{Efficiency change} = \frac{d_0^t(y_t, x_t)}{d_0^s(y_s, x_s)}$$

and

$$\text{Technical change} = \left[\frac{d_0^s(y_t, x_t)}{d_0^t(y_t, x_t)} \times \frac{d_0^s(y_s, x_s)}{d_0^t(y_s, x_s)}\right]^{1/2}.$$

To date, the most popular method to measure the distance functions $d_0^t(y_t, x_t)$, $d_0^s(y_s, x_s)$, $d_0^t(y_s, x_s)$ and $d_0^s(y_t, x_t)$, which make up Malmquist TFP index for linear models, has been the DEA-like linear programming methods suggested by Färe et al. (1994). Assuming a constant returns-to-scale (CRS) technology, this requires the solving of the following four linear programming problems:

$$\begin{aligned}
&\frac{1}{d_0^t(y_t, x_t)} = \max_{f,l} f, \\
&-f y_{it} + Y_t\, l \geq 0, \\
&x_{it} - X_t\, l \geq 0, \\
&l \geq 0,
\end{aligned} \tag{26}$$

$$\begin{aligned}
&\frac{1}{d_0^s(y_s, x_s)} = \max_{f,l} f, \\
&-f y_{is} + Y_s\, l \geq 0, \\
&x_{is} - X_s\, l \geq 0, \\
&l \geq 0,
\end{aligned} \tag{27}$$

$$\begin{aligned}
&\frac{1}{d_0^t(y_s, x_s)} = \max_{f,l} f, \\
&-f y_{is} + Y_t\, l \geq 0, \\
&x_{is} - X_t\, l \geq 0, \\
&l \geq 0,
\end{aligned} \tag{28}$$

and

$$\begin{aligned}
&\frac{1}{d_0^s(y_t, x_t)} = \max_{f,l} f, \\
&-f y_{it} + Y_s\, l \geq 0, \\
&x_{it} - X_s\, l \geq 0, \\
&l \geq 0,
\end{aligned} \tag{29}$$

The technical efficiency change may be decomposed into scale efficiency and "pure" technical efficiency components. To compare two production points, this requires the solution of two additional linear programmings (26) and (27), with convexity restriction $(\sum_{i=1}^{n} l_i = 1)$ added to each. For the case of n firms and T time periods, this would increase the number of LPs from $n(2T-2)$ to $n(4T-2)$ (Färe et al. (1994)).

The resulting expressions for efficiency measures EFF_i all rely upon the value of the unobservable U_i, obtained as the conditional expectation of U_i upon the observed value of $V_i - U_i$. Thus, the measure of technical efficiency relative to the stochastic production frontier is defined as:

$$EFF_i = E(Y_i^* | U_i, X_i) / E(Y_i^* | U_i = 0, X_i), \tag{30}$$

where Y_i^* is the production (or cost) of the i-th firm. The value of Y_i^* is taken to be equal to Y_i when the dependent variable is in the original units and will be equal to $\exp(Y_i)$ when the dependent variable is in logs. Moreover, the measure of cost efficiency relative to the cost frontier is also defined by (30). In the case of a production frontier, EFF_i will take a value between zero and one. In the case of a cost function, EFF_i will take a value between one and infinity. The resulting expressions for efficiency measures EFF_i (Jondrow et al. (1982) and Battese and Coelli (1988)) all rely upon the value of the unobservable U_i being predicted. The value is obtained by deriving expressions for the conditional expectation of these functions of the U_i, conditional upon the observed value of $V_i - U_i$ The relevant expressions for the production function case are provided in Battese and Coelli (1992, 1993, 1995), and the expressions for the cost efficiencies relative to a cost frontier have been obtained by minor alterations of the technical efficiency expressions in these papers.

5. Computer Software

To solve general problems of the least-squares estimation of econometric models and to solve problems in the calculation of price and quantity index numbers, the general-purpose econometrics package SHAZAM (White, 1993) is used. To conduct data envelopment analyses, (DEA) the Data Envelopment Analyses Program (DEAP Version 2.1) was developed (Coelli, 1996b). The principal options of the DEAP include: CRS and VRS DEA models that involve the calculation of technical and scale efficiencies, cost and allocative efficiencies; the Malmquist DEA methods for panel data to calculate indices of total factor productivity (TFP) change, technological change, technical efficiency change and scale efficiency change.

A computer program for stochastic frontier estimation, FRONTIER Version 4.1 (Coelli, 1996a) provides maximum-likelihood estimates of the parameters of a number of stochastic frontier production and cost functions. The models considered could accommodate panel data and assume firm effects that are distributed as truncated normal random variables. Estimates of standard errors are calculated along with individual and mean efficiency estimates. The FRONTIER Version 4.1 assumes a linear functional form. If one wants to estimate the most often-used non-linear Cobb-Douglas production function, one must log all o input and output data in order to convert the non-linear functional form to the linear. The FRONTIER Program can accommodate cross-sectional and panel data, time-varying and time-invariant inefficiency effects, cost and production functions, half-normal and truncated normal distributions, and functional forms with a dependent variable in logged or original units. Therefore, the FRONTIER Program may be used to estimate:

– a Cobb-Douglas production frontier using cross-sectional data and assuming a half-normal distribution;

– a translog production frontier using cross-sectional data and assuming a truncated normal distribution;

– a Cobb-Douglas cost frontier using cross-sectional data and assuming a half-normal distribution, and others.

Examples of the often-used Cobb-Douglas production functions are:

(i) The Cobb-Douglas production frontier:

$$\ln(Q_i) = b_0 + b_1 \ln(K_i) + b_2 \ln(L_i) + (V_i - U_i), \tag{31}$$

where Q_i, K_i and L_i are output, capital and labor, respectively, and V_i and U_i are assumed to be normally and half-normally distributed, respectively;

(ii) The Translog production frontier:

$$\ln(Q_i) = b_0 + b_1 \ln(K_i) + b_2 \ln(L_i) + b_3 \ln(K_i)^2 + b_4 \ln(L_i)^2 + \\ + b_5 \ln(K_i)\ln(L_i) + (V_i - U_i), \tag{32}$$

where Q_i, K_i, L_i and V_i are as defined earlier, and U_i has truncated normal distribution;

(iii) the Cobb-Douglas cost frontier:

$$\ln(C_i / W_i) = b_0 + b_1 \ln(Q_i) + b_2 \ln(R_i / W_i) + (V_i + U_i), \tag{33}$$

where C_i, Q_i, R_i and W_i are cost, output, capital price and labor price, respectively, and V_i and U_i are assumed to be normally and half-normally distributed, respectively.

References

Aigner, D. J., Lovell, C. A. K. and Schmidt, P. (1977), "Formulation and Estimation of Stochastic Frontier Production Function Models,“ *Journal of Econometrics*, 6, 21–37.

Ali, A. I. and Seiford, L. M. (1993), "Computational accuracy and infinitesimal in data envelopment analysis,“ *INFOR* 31: 290–297.

Appa, G. and Yue, M. (1999), "On setting scale efficient targets in DEA,“ *JORS* 50: 60–69.

Banber, R. D., Charnes, A. and Cooper, W. W. (1984), "Some Models for Estimation Technical and Scale Inefficiencies in Data Envelopment Analysis," *Management Science*, 30, 1078–1092.

Battese, G. E. and Coelli, T. J. (1988), "Prediction of Firm-Level Technical Efficiencies With a Generalised Frontier Production Function and Panel Data," *Journal of Econometrics*, 38, 387–399.

Battese, G.E. and Coelli, T.J. (1992), "Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India," *Journal of Productivity Analysis*, 3, 153–169.

Battese, G.E. and Coelli, T.J. (1993), "A Stochastic Frontier Production Function Incorporating a Model for Technical Inefficiency Effects," *Working Papers in Econometrics and Applied Statistics*, No. 69, Department of Econometrics, University of New England, Armidale.

Battese, G. E. and Coelli, T. J. (1995), "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data," *Empirical Economics*, 20, 325–332.

Battese, G. E., Coelli, T. J. and Colby, T. C. (1989), "Estimation of Frontier Production Functions and the Efficiencies of Indian Farms Using Panel Data From ICRISAT's Village Level Studies," *Journal of Quantitative Economics*, 5, 327–348.

Battese, G. E. and Corra, G. S. (1977), "Estimation of a Production Frontier Model: With Application to the Pastoral Zone of Eastern Australia," *Australian Journal of Agricultural Economics*, 21, 169–179.

Bauer, P. W. (1990), "Recent Developments in the Econometric Estimation of Frontiers," *Journal of Econometrics*, 46, 39–56.

Beyko, I., Bublyk, B. and Zinko, P. (1983), "Methods and Algorithms for Solving Optimization Problems," Naukova Dumka, Kyiv.

Charnes, A., Cooper, W. W., Lewin, A. Y., Morey, R. C. and Rousseau, J. (1985), "Sensivity and Stability Analysis in DEA," *Annals of Operations Research*, 2,139–156.

Coelli, T. J. (1992), "A Computer Program for Frontier Production Function Estimation: FRONTIER, Version 2.0," *Economics Letters* 39, 29–32.

Coelli, T. J. (1993), "Finite Sample Properties of Stochastic Frontier Estimators and Associated Test Statistics," *Working Papers in Econometrics and Applied Statistics,* No.70, Department of Econometrics, University of New England, Armidale.

Coelli, T. J. (1995), "Estimators and Hypothesis Tests for a Stochastic: A Monte Carlo Analysis," *Journal of Productivity Analysis*, 6, 247–268.

Coelli, T. J. (1996a), ”A Guide to FRONTIER Version 4.1: A Computer Program for Stochastic Frontier Production and Cost Function Estimation", *CEPA Working Paper 96/07*, Department of Econometrics, University of New England, Armidale.

Coelli, T. J. (1996b), ”A Guide to DEAP Version 2.1: A Data Envelopment Analysis (Computer) Program," *CEPA Working Paper 96/08*, Department of Econometrics, University of New England, Armidale.

Farell, M. J. (1957), ”The Measurement of Productive Efficiency,” *Journal of the Royal Statistical Society*, Series A, 120, 253–281.

Färe, R., Grosskopf, S. and Lovell, C. A. K. (1994), ”Production Frontiers,” Cambridge University Press, Cambridge.

Forsund, F. R., Lovell, C. A. K. and Schmidt, P. (1980), ”A Survey of Frontier Production Functions and of their Relationship to Efficiency Measurement,” *Journal of Econometrics*, 13, 5–25.

Greene, W. H. (1993), ”The Econometric Approach to Efficiency Analaysis,” in Fried H. O., Lovell C. A. K. and Schmidt S. S. (Eds), *The Measurement of Productive Efficiency*, Oxford University Press, New York, 68–119.

Himmelblau, D. M. (1972), ”Applied Non-Linear Programming,” McGraw- Hill, New York.

Hughes, M. D. (1988), ”A Stochastic Frontier Cost Function for Residential Child Care Provision,” *Journal of Applied Econometrics*, 3, 203–214.

Jondrow, J., Lovell, C. A. K, Materov, I. S. and Schmidt, P. (1982), ”On Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model,” *Journal of Econometrics*, 19, 233–238.

Kumbhakar, S.C., S. Ghosh and J.T. McGukin (1991), "A Generalized Production Frontier Approach for Estimating Determinants of Inefficiency in U.S. Dairy Farms," *Journal of Business and Economic Statistics,* 9, 279–286.

Lee, L. F. (1993), ”Asymptotic Distribution for the Maximum Likelihood Estimator for a Stochastic Frontier Function Model with a Singular Information Matrix,” *Econometric Theory*, 9, 413–430.

Meeusen, W. and van den Broeck, J. (1977), ”Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error,” *International Economic Review*, 18, 435–444.

Pitt, M. M. and Lee, L. F. (1981), ”Measurement and Sources of Technical Inefficiency in the Indonesian Weaving Industry,” *Journal of Development Economics*, 9, 43–64.

Reifschneider, D. and Stevenson, R. (1991), ”Systematic Departures from the Frontier: A Framework for the Analysis of Firm Inefficiency,” *International Economic Review*, 32, 715–723.

Seiford, L. M. and Thrall, R. M. (1990), ”Recent Development in DEA: The Mathematical Approach to Frontier Analysis,” *Journal of Econometrics*, 46, 7–38.

Schmidt, P. (1986), ”Frontier Production Functions,” *Econometric Reviews*, 4, 289–328.

Schmidt, P. and Lovell, C. A. K. (1979), ”Estimating Technical and Allocative Inefficiency Relative to Stochastic Production and Cost Frontiers,” *Journal of Econometrics*, 9, 343–366.

Stevenson, R. E. (1980), ”Likelihood Functions for Generalised Stochastic Frontier Estimation,” *Journal of Econometrics*, 13, 57– 66.

White, K. (1993), ”SHAZAM User's Reference Manual Version 7.0,” McGraw-Hill.

Yunos, J. M. and Hawdon D. (1997), “The efficiency of the National Electricity Board in Malaysia: An intercountry comparison using DEA,” *Energy Economics*, 19, 255–269.

Authors: Karl S. Althaler, Tatjana Slavova

Title: DEA Problems under Geometrical or Probability Uncertainties of Sample Data

Reihe Ökonomie / Economics Series 89

Editor: Robert M. Kunst (Econometrics)
Associate Editors: Walter Fisher (Macroeconomics), Klaus Ritzberger (Microeconomics)

ISSN: 1605-7996

Stumpergasse 56, A-1060 Vienna • ☎ +43 1 59991-0 • Fax +43 1 5970635 • http://www.ihs.ac.at

ISSN: 1605-7996